Semantic Interpretation of the map with Diabetes-Related Websites

Hongyi Shi, Marie-Christine Jaulent, Fabien Pfaender

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 160 (2019) 330–337

Procedia Computer Science

www.elsevier.com/locate/procedia

The 9th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2019)
November 4-7, 2019, Coimbra, Portugal

Semantic Interpretation of the map with Diabetes-Related Websites

Hongyi Shi[a,*], Marie-Christine Jaulent[a], Fabien Pfaender[b]

[a]*Inserm, Sorbonne Université, Université Paris 13, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé, LIMICS, F-75011 Paris, France*

[b]*UTSEUS, Shanghai University, Shanghai, China; Costech EA2223, Université de Technologie de Compiègne, Compiègne, France*

Abstract

Diabetes as a chronic disease requires continuous medical care and constant patient self-management which involve several stakeholders to improve health outcome and patient quality of life. In our prior work, we used the networks of World Wide Web to highlight how stakeholders of diabetes link to each other online. The aim of this study is to use a semantic approach focusing on the diabetes-related websites to better understand the common interest shared by the same clusters which were detected in our previous study of stakeholders on diabetes. To achieve this, we employed the data annotation and machine learning to study which combinations of tags can predict or explain the clusters. In the end, a total of 430 websites which are detected into 5 clusters have been tagged with 38 different tags from 6 different dimensions in this study. Although the result shows a very low prediction performance using tags to determine the clusters of diabetes-related websites, except for cluster 1 and cluster 2, this reflects the community reality: a mix of websites of different types that create a mixed but localized space. It proves the community can have a tagging scheme occasionally but it is still hard to use semantical approach to predict accurately the clusters.

Keywords: diabetes mellitus; information networks; data visualization; clustering; semantic interpretation; machine learning

* Corresponding author. Tel.: +3-361-748-1849.
E-mail address: maryshi@live.cn

1877-0509
10.1016/j.procs.2019.11.083

1. Introduction

Diabetes is one of the largest global health emergencies of the 21st century. According to International Diabetes Federation Diabetes Atlas, 8th edition, there were 425 million people with diabetes until 2017 and this figure could rise to 642 million by 2040 [1]. Diabetes can be treated and its consequences could be avoided or delayed with a specific diet, physical activity, medication and regular screening and treatment for complications [2]. In the field, we can establish a preliminary list of possible key stakeholders: hospitals, physicians, research laboratories, pharmaceutical companies, medical and social relay structure, associations, health related NGOs (non-governmental organizations), patients and their families around diabetes care. All of these key diabetes actors play their own roles in the health care system but the nature of their relationship, if any, remains unexplored outside of a patient-centric paradigm. The web, invented at the beginning of the 1990's, has gradually became a major platform itself hosting billions of resources accessible from a URL (Uniform Resource Locator) [3]. The web is also the source of scientific investigation in itself and the network structure of a hyperlinked environment can be a rich source of information for whom looking for an effective means to collect, analyze and understand it [4,5]. Web structures are proxy to human organizational structures [6]. Thus, we propose to explore by map and analyze the hyperlink structure of diabetes stakeholders online. As a matter of fact, the web contains crucial aspects of the embodiment of social factors: personal blogs, institutional websites, health-oriented media, etc.

In a previous work, we already created and explored a network with 430 diabetes-related websites included [7]. With combination of state-of-art tools, Hyphe (a web crawler) and Gephi (an open source software for network visualization and analysis) to crawl and visualize topic-sensitive networks to analyze chronic diseases stakeholders' organizations and communities. We applied a community detection algorithm aimed at detecting clusters of similarly connected nodes [8]. It does so by maximizing the modularity quality metric in all possible partitions. Modularity measures the difference between the edge density in the partition and a randomized graph with the same number of nodes and the same degree distribution [9]. The result showed that diabetes online community has its own space and is organized with clusters and distribution of websites of relative importance. Indeed, five clusters have been discovered and can be visualized in figure 1 using different colors. In addition, we used a degree centrality algorithm [10] which assigns an importance score to each website based on the number of links held by each node to get several central websites in our 430 diabetes-related websites. Node Degree reveals the number of hyperlinks which are from 2 to 300 related to one web entity.

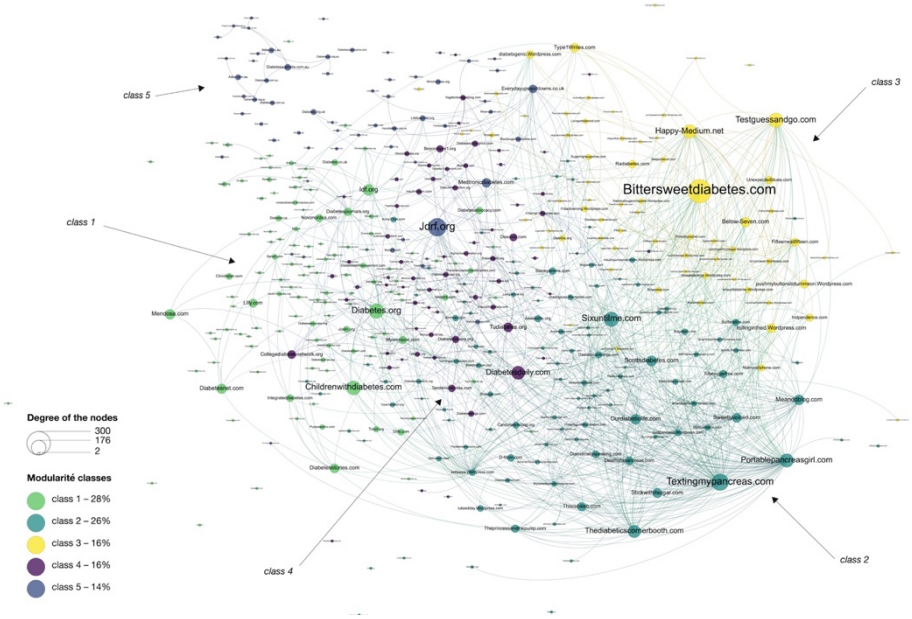


Fig. 1. Diabetes hyperlink structure associated to 5 discovered communities.

However, what these clusters refer to? How to semantically explain the relationship inside the network of diabetes-related websites to explain why some websites are getting closer than others? What do they have in common if they are in the same cluster which was detected by a community detection algorithm?

The hypothesis of this study is that a semantic description of the diabetes-related websites with tags could allow to predict the community clusters of the network. The overarching purpose of this study was to investigate if there is a specific reason, in relation with some semantics, that explains why a website belongs to one class rather than another in the map obtained from our previous work. To achieve that, we divided this exploratory study into three steps: (1) to define and organize the best tags describing the semantic content of websites; (2) to define and set an annotation process to annotate the 430 websites manually using the tags; (3) to apply various machine learning methods to predict the class of a website from the annotations.

2. Methods

2.1. Material

The material is a dataset including the 430 diabetes-related websites that were obtained in the previous study using a web crawler Hyphe. Each website is associated with the class it belongs to. Table 1 shows a sample of 10 websites with their belonging clusters from the raw dataset. Here we use class 1, class 2, class 3, class 4 and class 5 as notations to distinguish the classes.

Table 1. Sample of the dataset with 10 random websites and their belonging clusters.

Website	Description	Class
Diabeticinvestor.com	Diabetes Investor is the premier subscription-based content publisher that provides real time analysis of the business of diabetes.	1
Affordableinsulinproject.org	The Affordable Insulin Project offers tools, resources, and data so that people impacted by today's rising health care costs can positively influence the affordable access to this life essential drug.	1
Smashtastic.Wordpress.com	Smashtasitic is the blog of one 27 years old girl with type1 diabetes.	2
Sweetlyvoiced.com	Sweetlyvoiced is a blog of one mom with type1 diabetes.	2
Stripsafely.com	StripSafely is a Diabetes Online Community (DOC) to help the general public understand that there are inaccurate blood glucose test strips and meters on the market.	3
Thetype2Experience.com	The Type 2 Experience is a collaboration blog created by a group of friends who live with type 2 at different levels and with different backgrounds.	3
Adorndesigns.com	The Adorndesigns is the online shop providing "High Style – Low Profile" handbags for diabetics.	4
Dexcom.com	Dexcom is a global Continuous Glucose Monitoring (CGM) System company.	4
Childrenwithtype1Diabetes.org	Children With Type 1 Diabetes is an organization to provide support and information to parents of children diagnosed with Type 1 Diabetes.	5
Jdrf.org	JDRF is the leading global organization funding type 1 diabetes research.	5

2.2. Data annotation

We employed inductive thematic analysis which is one of the qualitative analytic method [11] to construct categories to describe the content of the websites with different dimensions. One diabetes expert with 15 years type 1 diabetes experience proposed the initial set of categories to annotate the websites according to the stakeholders of diabetes, language and type of the diabetes-related websites, type of diabetes and diabetes-related topics, etc. Then, the project team reviewed part of websites randomly and annotate each website using this set of categories. From the two propositions, we set categories with their own possible values to decide the final tags for the annotation process. The

diabetes expert annotated manually the 430 websites with the final tags. It took around one month to complete the whole annotation procedure.

2.3. Class prediction

To study which combinations of tags can predict or explain the classes/communities/clusters according to our previous study, we used RapidMiner studio framework [12] to apply 7 different state-of-the-art clustering models to our dataset. The RapidMiner Studio is a visual workflow designer software to import, prepare, and clean the data and then perform state of the art data science and machine learning algorithm. The modeling methods are: *Naive Bayes*, *Generalized Linear Model*, *Deep Learning*, *Decision Tree*, *Random Forest*, *Gradient Boosted Trees* and *Support Vector Machine*. For testing purposes, the dataset was split into a training set for the machine to learn and a testing set to determine the quality of the prediction. We used 38 tags as features and the cluster ID as a label to feed different models. We chose a set of 60% random websites for training and 40% for testing and repeated several times the experiment. This ratio for training/testing could be higher but with the risk of overfitting the data.

3. Results

3.1. Category sample from Inductive Thematic Analysis

Six categories with different values are identified: Status of stakeholders, Language of websites, Type of websites, Organizations, Type of diabetes, and Diabetes-related topics. The 6 categories providing 38 values are presented in Table 2. For some of them, the values are mutually exclusive as status, language, type of websites, organizations and for the others multiple values are authorized as type of diabetes and diabetes-related topics.

Table 2. The output of 6 categories with 38 values for tagging 430 diabetes-related websites.

Status of Stakeholders	Language of Websites	Type of Websites	Organizations	Type of Diabetes	Diabetes-Related Topics	Total
Non-Profit	English	Portal	Individual	Type 1	Prevention	
Profit	Multilingual	Information	Association	Type 2	Treatment	
		Blog	Society	Gestational	Self-management	
		Forum	Federation	Pre-diabetes	Advocacy	
		E-commerce	Charity		Complications	
		Click-to-donate	Company		Psychological Support	
			Program		Accessories	
			Conference		Sport	
			Hospital		Diabulimia	
			Clinic			
			Pharmacy			
			Laboratory			
			Consulting			
			Media			
			Online Community			
2	2	6	15	4	9	38
Maximum number of possible tags from each category to annotate one web-site						
1	1	1	1	4	9	17

3.2. Annotation process and tags distribution

Using the set of tags, the whole corpus (430 websites) have been tagged with 36 different tags values instead of 38. Indeed, the 4 tags “profit”, “non-profit” (Status of Stakeholders) “English” and “Multilingual” (Language of Websites) can be reduce to 2 tags: “Profit” (if the tag is absent, it means that it is a non-profit website) and “English” (if this tag is absent, it means that it is a multilingual website). The 36 tags are unequally spread among the websites. As shown on figure 2, some tags are almost ever present while others are very specific and almost never used. However, each tag has been at least tagged once and no tag covers the whole corpus. On average, a tag is used 80.8 times but the count follows a power law distribution and tags like diabetes type 1, language (English), self-management and website type information are widely found in more than half of the website (count > 215). They describe the corpus as a whole and define the broader topic this corpus is about: diabetes type 1, in English providing information about how to manage one’s diabetes. In the other hand, 20 tags are used in less than 10% of the websites and are either very specific to a niche-like topic as diabulimia which is one food disorder related to diabetes (Diabulimiahelpline.org) or simply non-relevant at all, for example, some organization type like consulting (Diabeticinvestor.com) which is giving the information to people who need to develop their business related to diabetes or pharmacy (Diabetesexpress.ca).

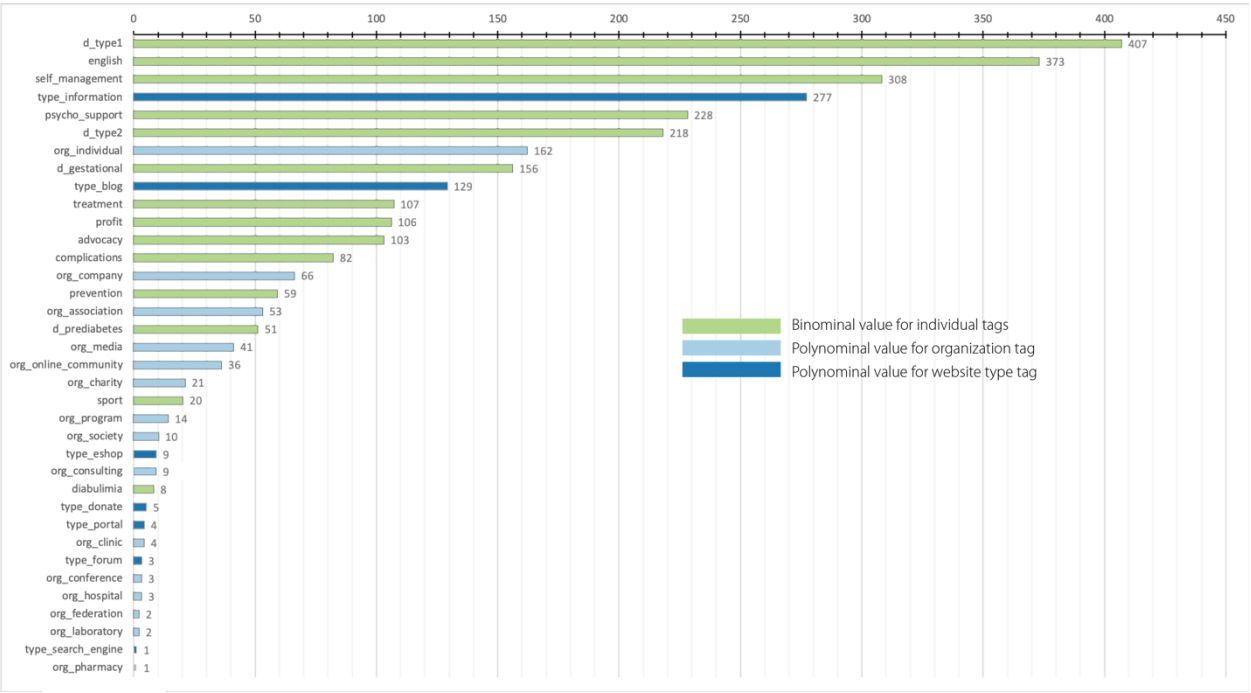
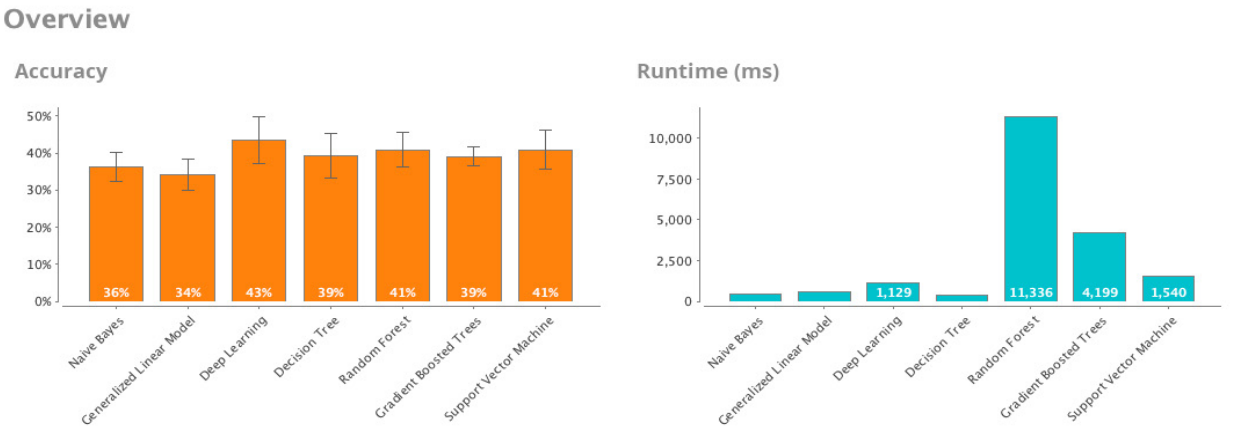


Fig. 2. Tags count on the 430 websites (The tag names are prefixed with their abbreviated category)



Accuracy

Model	Accuracy	Standard Deviation	Runtime
Naive Bayes	36.2%	± 4.0%	474 ms
Generalized Linear Model	34.2%	± 4.2%	572 ms
Deep Learning	43.5%	± 6.4%	1 s
Decision Tree	39.3%	± 6.1%	376 ms
Random Forest	40.9%	± 4.7%	11 s
Gradient Boosted Trees	39.1%	± 2.6%	4 s
Support Vector Machine	40.9%	± 5.4%	2 s

Fig. 3. The global performances of 7 models and runtime to predict clusters from tags.

3.3. Class prediction

The seven machine learning models were applied to the data obtained by the annotation process to predict clusters according to tags. The global performances and runtime to predict clusters from tags are present in figure 3. The performance refers how many times machine can predict correctly the website's class from the tags model. Precision is defined as the number of true positives over the number of positives plus the number of false positives. Recall is defined as the number of true positives over the number of true positives plus the number of false negatives. A system with high recall but low precision returns many results, but most of its predicted labels are incorrect when compared to the training labels. A system with high precision but low recall is just the opposite, returning very few results, but most of its predicted labels are correct when compared to the training labels. An ideal system with high precision and high recall will return many results, with all results labelled correctly. However, no model stands out as a good predictor of our dataset.

The accuracy of the 7 models is 40% at best. This indicates a weak prediction capability. Indeed, no model really stand out indicating that the cause of the weakness shall probably lie in the data as all the models use the same data as an entry point. To get into more details about the performance of the model, we chose the relatively high accuracy model, random forest which generate a forest of decision trees of variable size and depth. The optimal parameter for the random forest is 140 trees with a maximal depth of two, which is an average parameter setting for such a dataset property (in terms of data count and data structure). The detail accuracy of the random forest model on figure 5 shows two distinct phenomena. On one hand, the model can predict something when websites belong to cluster 2 or cluster 1 (the two most populated clusters) even if its accuracy in doing so is very low. Actually, the class recall of class 1 is 80% while the class recall of class 2 is almost 70%, indicating that the model returns actual predictions for these classes; however, their low-class precision indicates that even though we can predict something at all for them, making

a true prediction is still hard to do with 42% precision average, (35.94+48.33)/2 (see figure 4). To summarize, many results are returned but most of the prediction are incorrect. On the other hand, this is not true at all for the 3 other classes. As a matter of fact, none of the 3 other clusters (class 3,4,5) have been predicted by the random forest model. This indicate that the tags we created worked on two clusters only but are not specific enough to really predict these two clusters while the 3 remaining clusters are simply not represented by the tags. The latter are indeed too heterogeneous in their tags' distribution and not specific enough for a group of tags to be predicted accurately.

accuracy: 41.93% +/- 4.00% (micro average: 41.94%)

	true class2	true class4	true class1	true class3	true class5	class precision
pred. class2	23	7	7	20	7	35.94%
pred. class4	0	0	0	0	0	0.00%
pred. class1	10	9	29	3	9	48.33%
pred. class3	0	0	0	0	0	0.00%
pred. class5	0	0	0	0	0	0.00%
class recall	69.70%	0.00%	80.56%	0.00%	0.00%	

Fig. 4. Random forest model performance for predicting each cluster according to the tags.

4. Discussion

Results show a low prediction performance by using tags to provide a semantic explanation of the clusters of diabetes-related websites obtained in our previous work. While looking at the tags distribution, this result reflects the fact that some tags are specific to cluster 1 or 2 and they are the ones with maximum weight found from 7 methods. This means that two of our clusters do have an identity that is relatively easy to predict with tags. However, even for cluster 1 or 2, it is hard to explain the prediction by a specific combination of the tags (a semantic signature).

In this paper, we used a semantic approach based on tags chosen by one expert. After the annotation, we found some tags are too specific and unnecessarily complicate the annotation process. They are maybe useful from a user point of view but are not good enough to become the discriminating mechanism to identify regions or localities inside the corpus. For example, tags as clinic and hospital can be replaced by healthcare facilities to just simplify the tags. One perspective will be to use alternative approaches, like natural language processing, to automatically extract the most relevant tags from the whole set of websites to improve the prediction rate. Another approach could be to learn the most appropriate tags from the set of websites for each cluster.

In addition, we found that the main two clusters with the best prediction rate are also containing the largest number of websites. Indeed, the resolution limitation of modularity makes the algorithms unable to detect small communities [13], even for one of the best model in our study which is Random Forest. This leads us to a new hypothesis that the current network is not big enough for machine learning to predict the right cluster. In future work, using the crawling procedure, we intend to extend the number of websites to above 5000, with the objective to predict more accurately the related clusters. It will also help us to get the more appropriate size of each cluster for learning methods. However, this will raise the big issue of time consuming for the manual annotation process. Indeed, considering 430 websites took almost 1 month to tag so that tagging 5000+ websites seems to be a too time-consuming job. One solution to this is to again use Nature Language Processing (NLP) solutions for the automatization of the annotation process. If we can use NLP to get accurate tags, can we also use it to facilitate the annotation process? Somehow, the most difficult part for teaching machine to do the annotation is not only to detect the presence of a tag in the web page but rather to understand the meaning behind. When experts annotate manually, usually the first step is to read "About Us" or "Who

we are" to get the general idea about the website. For example, like www.jdrf.org, in "About Us" part, it is clearly written "JDRF is the leading global organization funding type 1 diabetes (T1D) research." And their mission is "Improving lives today and tomorrow by accelerating life-changing breakthroughs to cure, prevent and treat T1D and its complications." [14] So it is easier to find the tags for Non-profit, Type1, Association, Prevention, Treatment and Complications. But with some other websites, it takes time to read the webpages to get the main information to annotate. Also with the different ways to express the same meaning, like type1 diabetes, T1D, insulin dependent diabetes, Juvenile diabetes, we need to collect the enough words in our corpus to describe the contents.

5. Conclusion

Although this study doesn't conclude that all clusters can be explained at the semantic level, at least two clusters are clearly defined by a few specific tags and the others are mixed. Without the successful predicting performances, we can better understand the tags distributions and their relative importance. This can help us to refine the tags for future broader analysis of the diabetes web space. The result of the clusters prediction also reflects the community reality: a mix of websites of different types that create a mixed but localized space. It proves the community has a tagging scheme sometimes but it is still hard to use semantical approach to predict accurately the clusters. However, diabetes online space does exist and this new virtual world as an open resource should be known by everyone. In practice, this work will allow us to improve the search engines to find more relevant and accurate information by using semantic tags. Our work as an entry point presented how to combine the network visualization and community detection to do the network analysis domain in diabetes. This new approach can be applied for network analysis in any other diseases domain.

Acknowledgements

This paper was made possible by the cooperation program by Sorbonne Université and Chinese Scholarship Council. Chinese Scholarship Council offers fellowship during I'm studying PhD in Sorbonne Université. I would like to thank LIMICS, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé, UMRS_1142 for the support and this work is also supported partly by Complex City Laboratory, Shanghai.

References

[1] Ogurtsova K, Rocha Fernandes JD, Huang Y, Linnenkamp U, Guariguata L, Cho NH, Cavan D, Shaw JE, Makaroffad LE. (2017) "IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040." *Diabetes Research and Clinical Practice* **128**:40-50.

[2] Fadupin GT, Keshinro OO. (2001) "Factors influencincing dietary compliance and glycaemic control in adult diabetic patients in Nigeria." *Diabetes Int.* **11**:59–61.

[3] Miller J. (2016) "What is the difference between the World Wide Web and the Internet?" https://blog.telegeography.com/whats-the-difference-between-the-world-wide-web-and-the-internet. Archived at: http://www.webcitation.org/77PMSWTUA.

[4] Kleinberg JM. (1999) "Authoritative sources in a hyperlinked environment." *Journal of the ACM (JACM)* **46(5)**:604-632.

[5] Pfaender F, Jacomy M, Fouetillou G. (2006) "Two visions of the web: from globality to localities." *ICTTA* **1**:566-571.

[6] Pirolli P, Pitkow J, Rao R. (1996) "Silk from a sow's ear: extracting usable structures from the Web." *ACM Press* **33**:118-125.

[7] Shi H, Pfaender F, Jaulent MC. (2019) "Mapping the hyperlink structure of diabetes online communities." Proceedings of the 17th World Congress of Medical and Health Informatics; Lyon, France.

[8] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre R. (2008) "Fast unfolding of communities in large networks." *Journal of Statistical Mechanics: Theory and Experiment* **10**: P10008.

[9] Mingming C, Konstantin K, Szymanski BK. [2014] "Community detection via maximization of modularity and its variants." *IEEE* **1(1)**: 46-65.

[10] Everett MG, Borgatti SP, Carrington PJ, Scott J, Wasserman S. (2005) "Models and methods in social network analysis." *New York: Cambridge University Press*: 57-76.

[11] Braun V, Clarke V. (2006) "Using thematic analysis in psychology." *Qualitative research in psychology* **3(2)**:77-101.

[12] Gupta G, Malhotra S. (2015) "Text documents tokenization for word frequency count using rapid miner (taking resume as an example)" *International Journal of Computer Application.* Tool from: https://rapidminer.com

[13] Lancichinetti A, Fortunato S. (2011) "Limits of modularity maximization in community detection." *Physical Review E* **84(6)**:066122.

[14] https://www.jdrf.org/about. Archived at: http://www.webcitation.org/77PNGOZxs